#2 Supervised Learning

by Saturdays.AI

Saturdays.AI

Machine Learning

Saturdays.AI
Donostia

Schedule

State of the course

Session 2 Review

Challenge

Notebook + resources

State of the course

Saturdays.AI
Donostia

#1 Cleaning & Exploratory Data Analysis ✔️

#2 Supervised Learning 🔴 Today!

#3 Decision Trees & Random Forest ➡️ SOON

#4 Unsupervised Learning + Clustering ➡️ SOON

#5 Time Series Analysis + Data Viz ➡️ SOON

#6 Neural Networks, Gradient Descent ➡️ SOON

#7 NLP ➡️ SOON

Little general view...

Saturdays.AI
Donostia

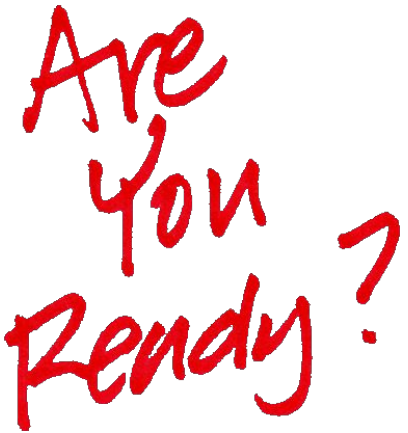
Little general view... CRISP-DM

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	**Select Data** *Rationale for Inclusion/ Exclusion*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Clean Data** *Data Cleaning Report*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Build Model** *Parameter Settings* *Models* *Model Descriptions*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Integrate Data** *Merged Data*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Format Data *Reformatted Data*			
		Dataset *Dataset Description*			

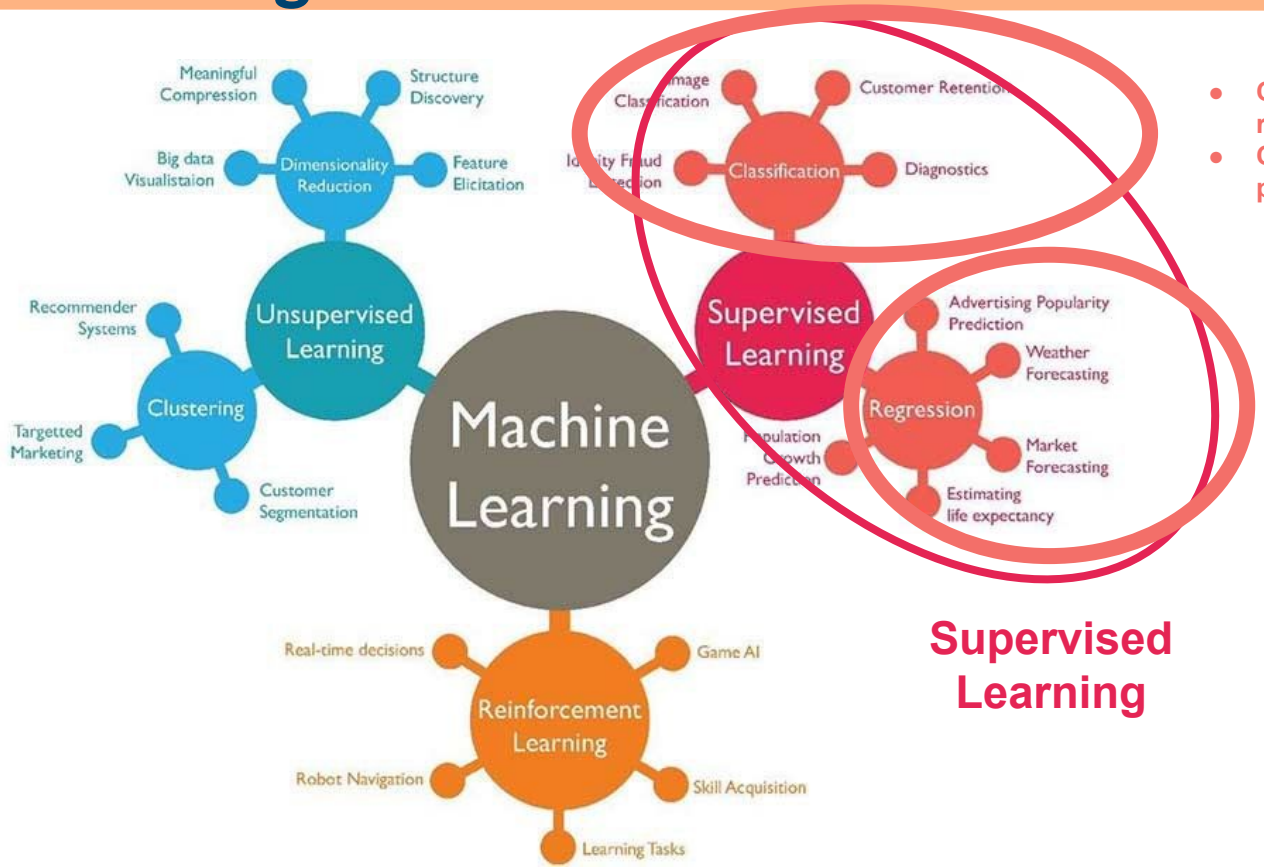
Regression & Model Validation

Saturdays.AI
Donostia

- General model topics
  - Overfitting/Underfitting
  - Bias/Variance
  - Cross-validation
  - Regularization
  - Model evaluation
- Models: Linear Models, Naive Bayes, AR, SVM



Machine learning

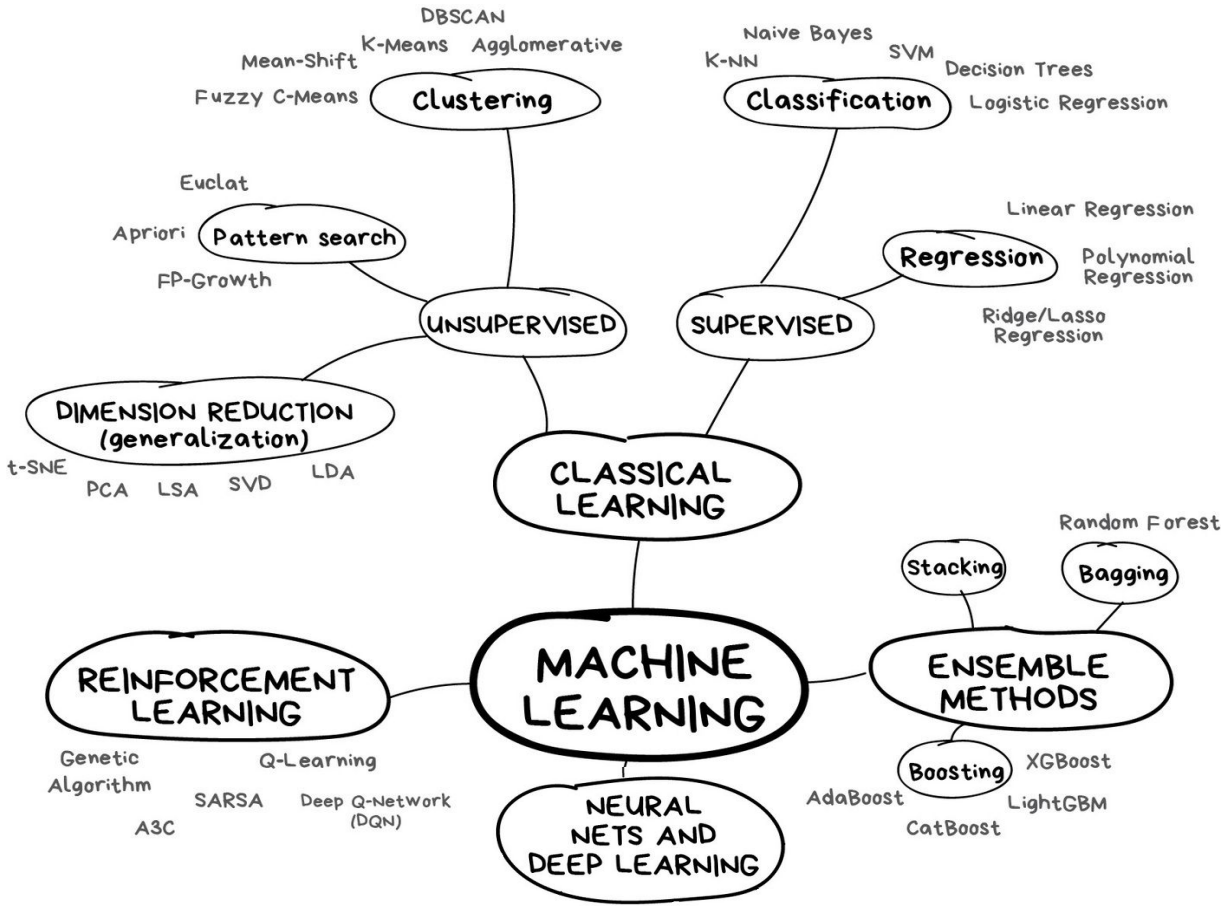


- Categorical response variable
- Classification problems

- Linear response variables
- Linear problems

Supervised Learning

ML Algorithms



Supervised learning: linear problem (Ejemplo)

Saturdays.AI
Donostia

HOUSE PRICE Dataset

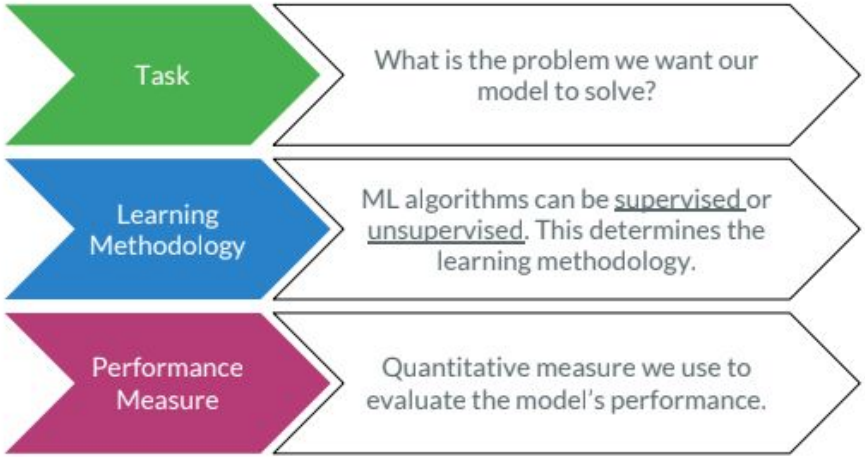
Id	SalePrice	MSZoning	YearBuilt	LotArea	Street
1	208500	RL	2003	8450	Pave
2	181500	RL	1976	9600	Pave
3	223500	RL	2001	11250	Pave
4	140000	RL	1915	9550	Pave
5	250000	RL	2000	14260	Pave
6	143000	RL	1993	14115	Pave
7	307000	RL	2004	10084	Pave
8	200000	RL	1973	10382	Pave
9	129900	RM	1931	6120	Pave

Regression & Model Validation

Saturdays.AI
Donostia

Modeling Phase

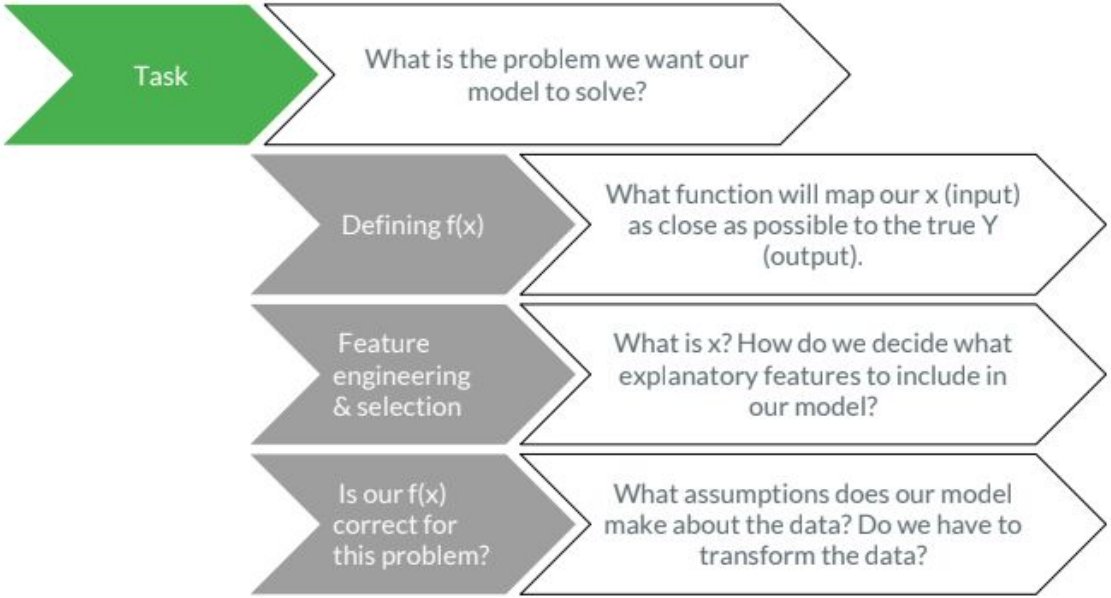
All models have 3 key components: a task, a learning methodology and a performance measure



Regression & Model Validation

Saturdays.AI
Donostia

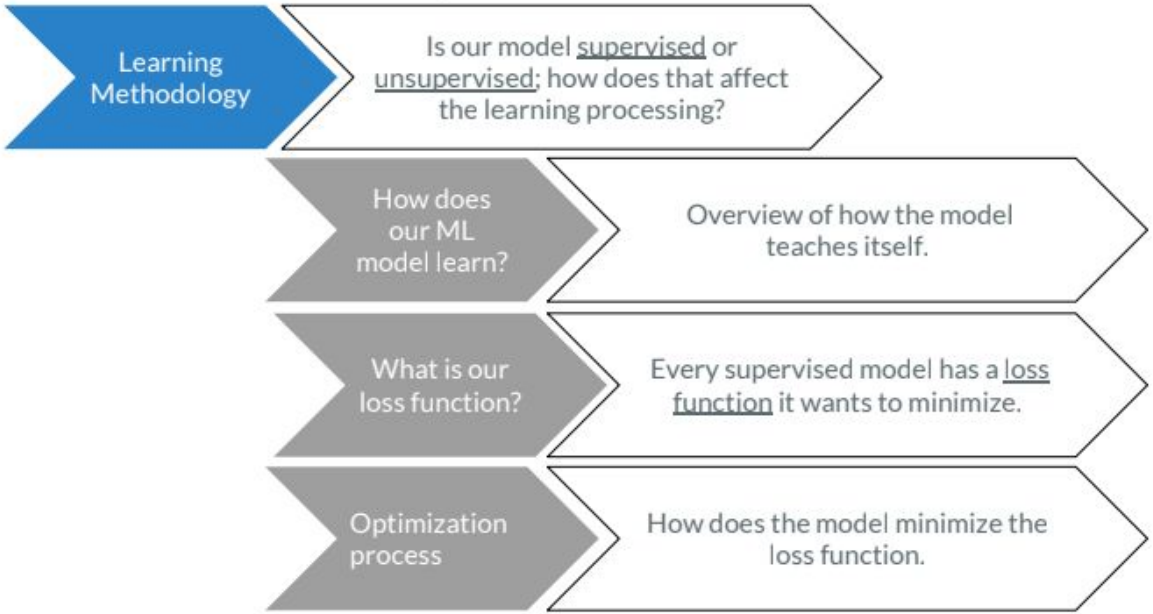
Today we are looking closer at each component of the framework:



Regression & Model Validation

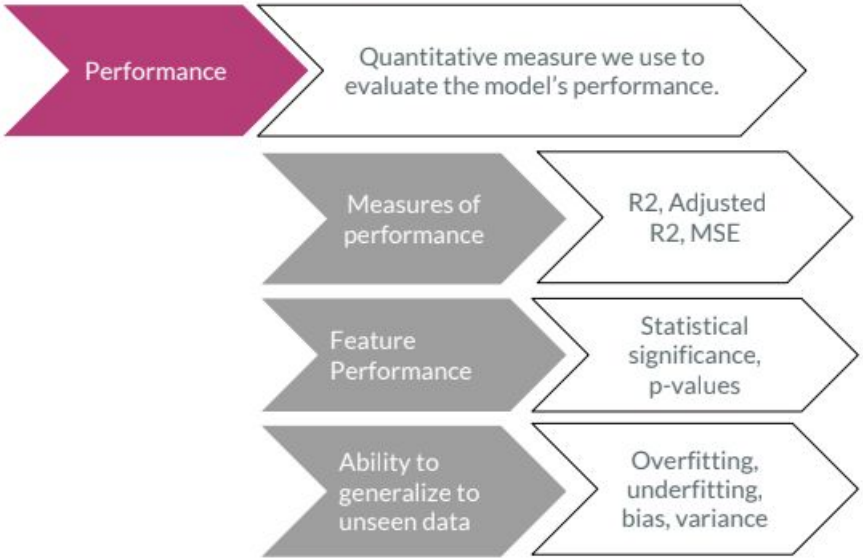
Saturdays.AI
Donostia

Learning methodology: how does the model **learn** the function that best maps x to the true Y?



Regression & Model Validation

Performance: How do we evaluate how useful the mode is, and how we can improve it?



Saturdays.AI
Donostia

Regression Loss Function

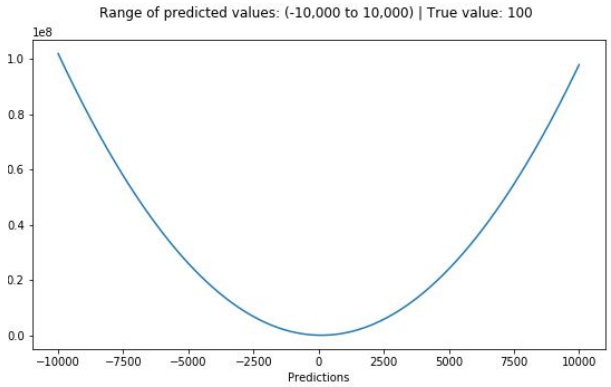
Saturdays.AI Donostia

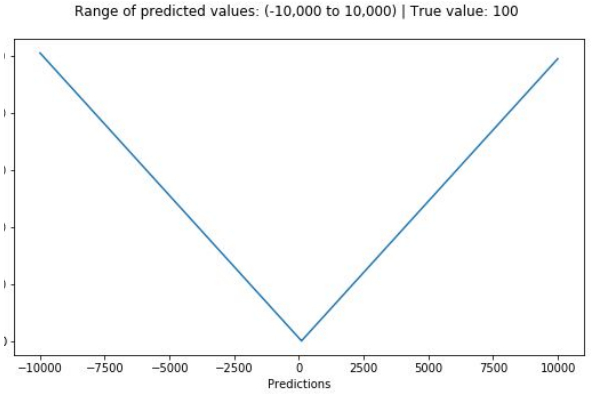
- Regression
  - Mean Square Error/ Quadratic Loss
  - Mean Absolute Error
  - Huber Loss/ Smooth Mean Absolute Error
  - Log cosh Loss
  - Quantile Loss

Q: ¿Cómo de bien o mal vamos en el entrenamiento?

$$MSE = \frac{\sum_{i=1}^{n} (y_i - y_i^p)^2}{n}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_i^p|}{n}$$





Regression Loss Function

Saturdays.AI Donostia

- Regression
  - Mean Square Error/ Quadratic Loss
  - Mean Absolute Error
  - Huber Loss/ Smooth Mean Absolute Error
  - Log cosh Loss
  - Quantile Loss

MAE vs. RMSE for cases with slight variance in data

ID	Error	|Error|	Error²
1	0	0	0
2	1	1	1
3	-2	2	4
4	-0.5	0.5	0.25
5	1.5	1.5	2.25

MAE: 1 RMSE: 1.22

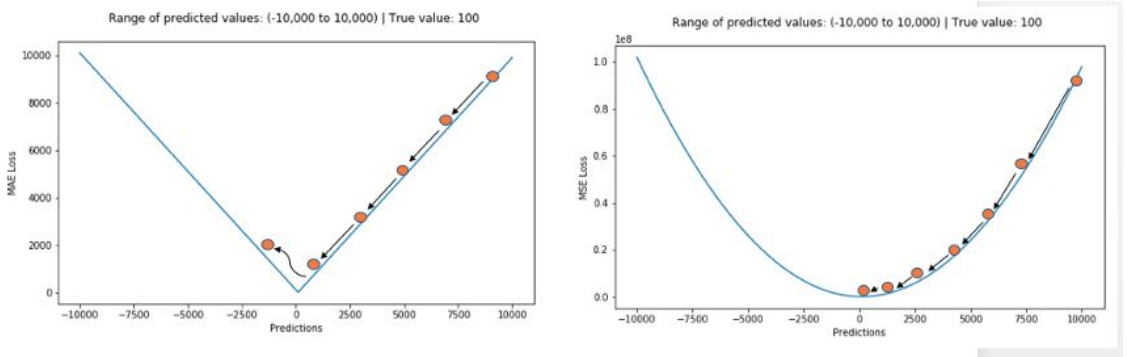
MAE vs. RMSE for cases with outliers in data

ID	Error	|Error|	Error²
1	0	0	0
2	1	1	1
3	1	1	1
4	-2	2	4
5	15	15	225

outlier

MAE: 3.8 RMSE: 6.79

En la práctica, esto sucede en la computadora (cuando aprende estamos entrenando el modelo)...



Measures of performance

Saturdays.AI
Donostia

Todo parte de la MATRIZ DE CONFUSIÓN...

Modelo

Realidad

	PREDIJO SI	PREDIJO NO
REAL SI	80	10
REAL NO	10	100

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

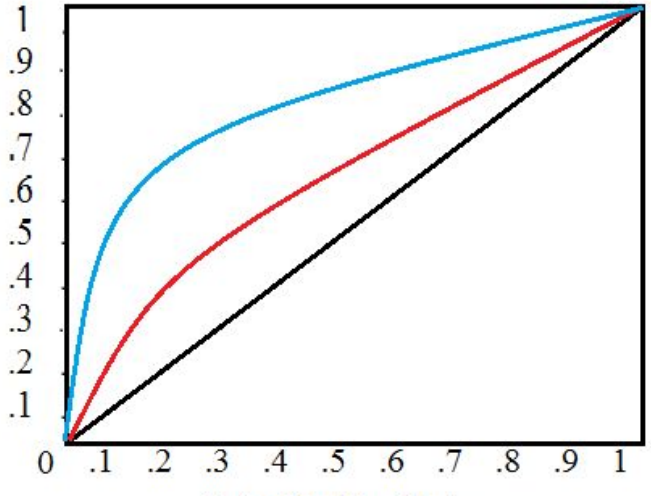
$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

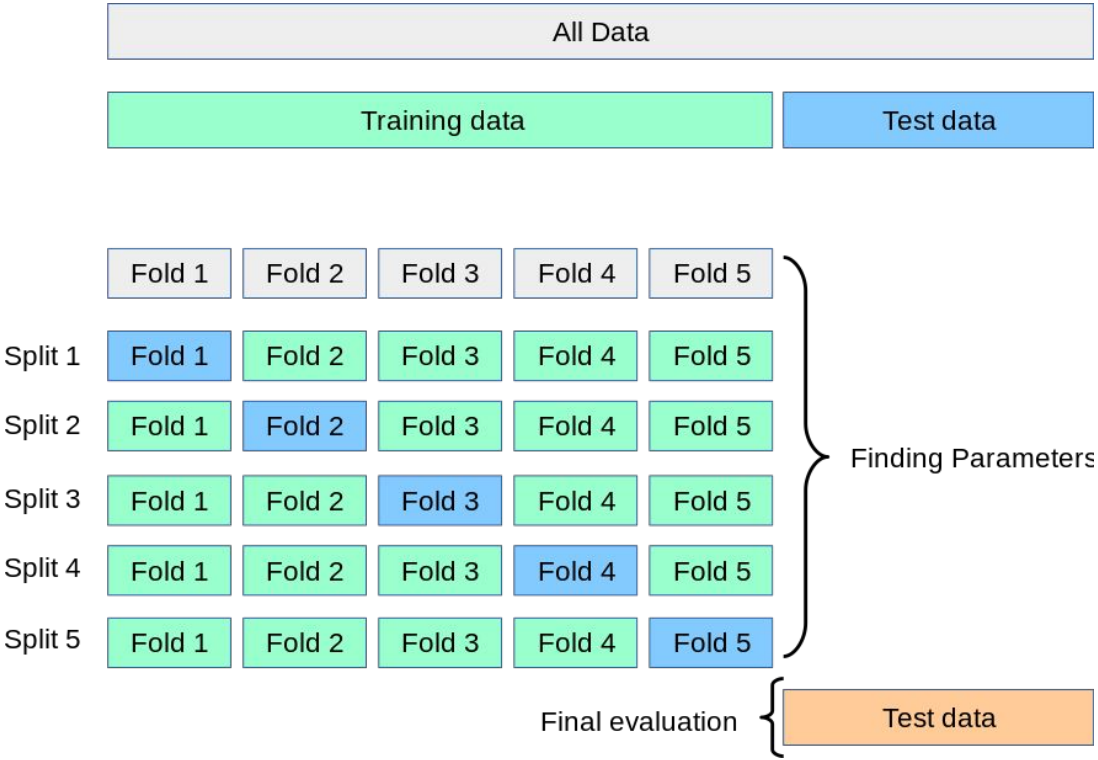
Measures of performance

ROC curve → Plots the true positive rate (TPR) versus the false positive rate (FPR) as a function of the model's **threshold** for classifying a positive.

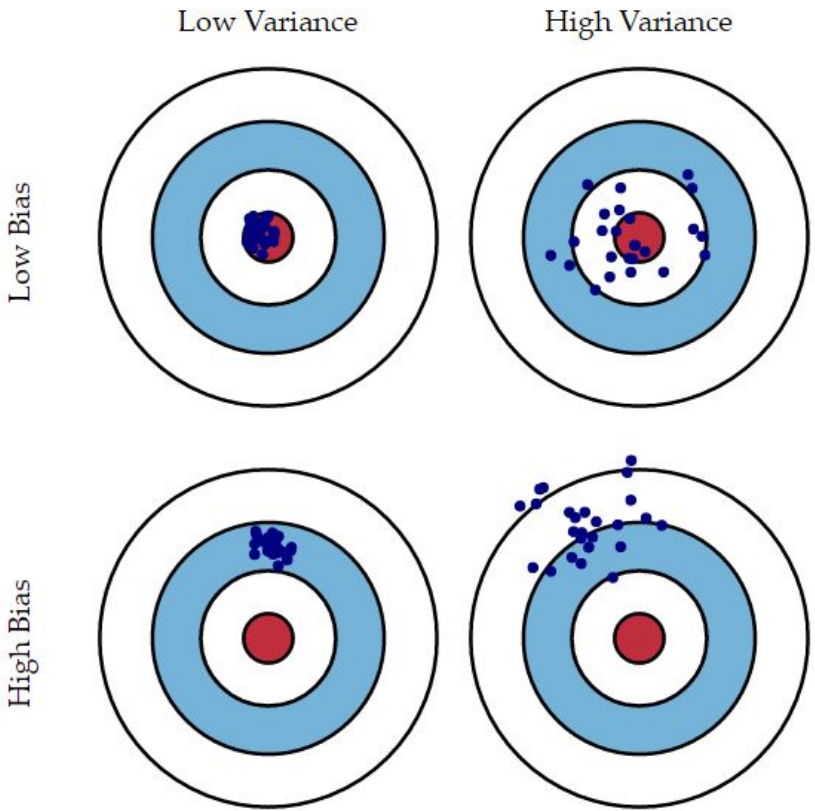
AUC → Metric to calculate the **overall performance** of a classification model based on area under the ROC curve.



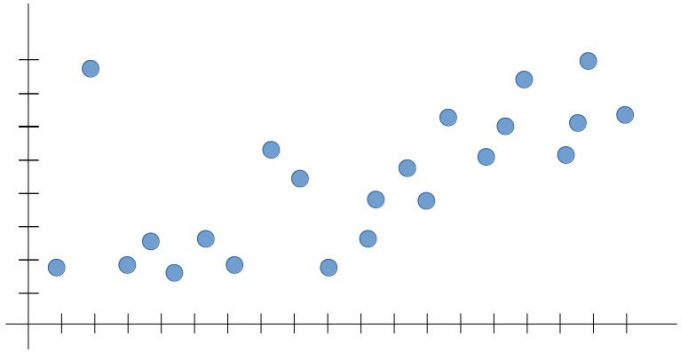
Cross-validation

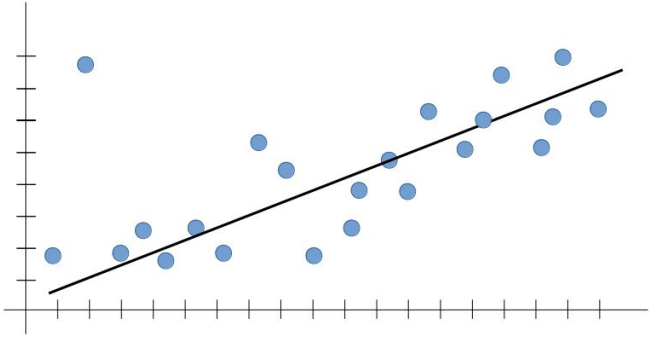


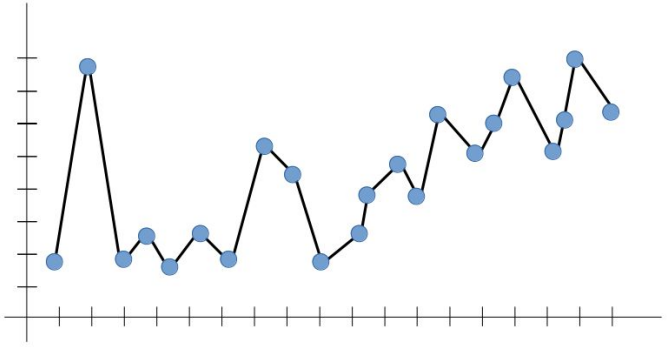
Bias vs. Variance

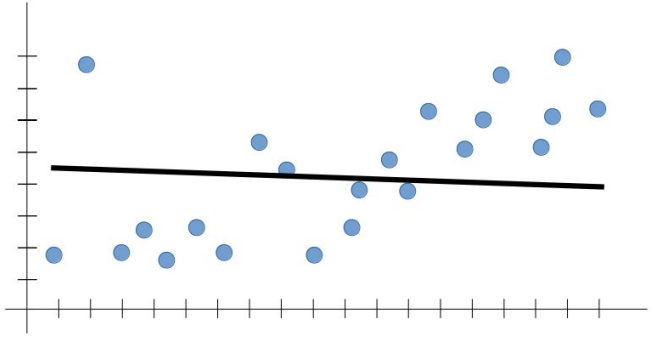


Underfitting vs. Overfitting

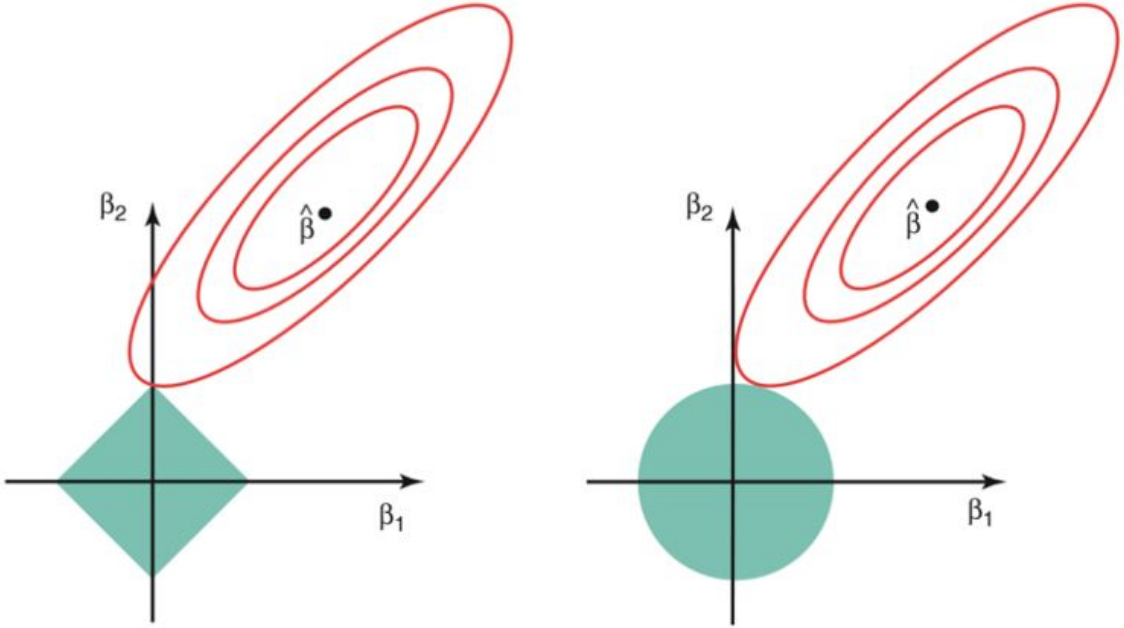






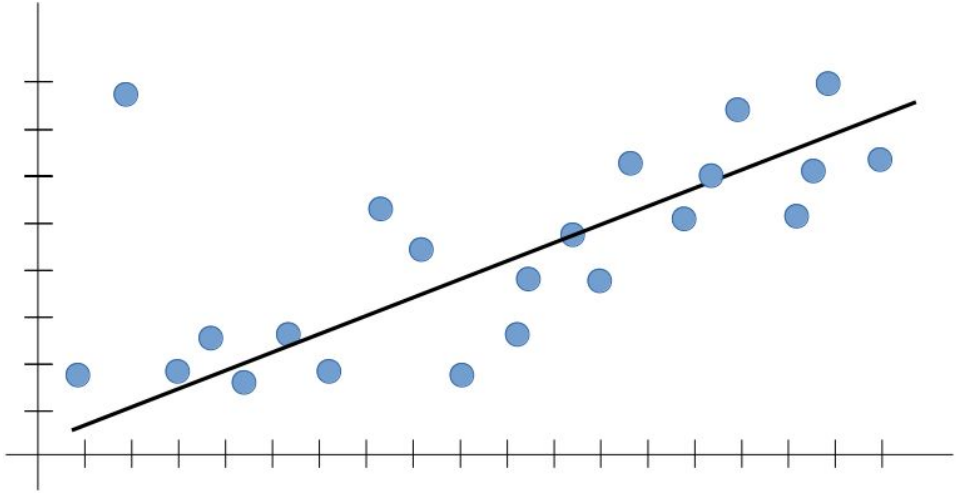


Regularization



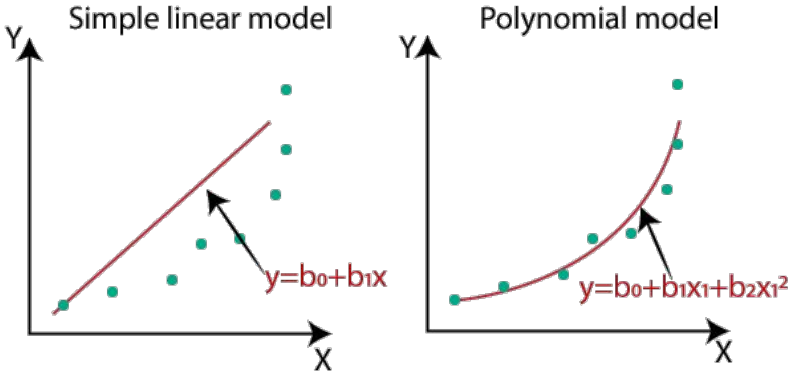

Let’s dive into models!

Linear Regression



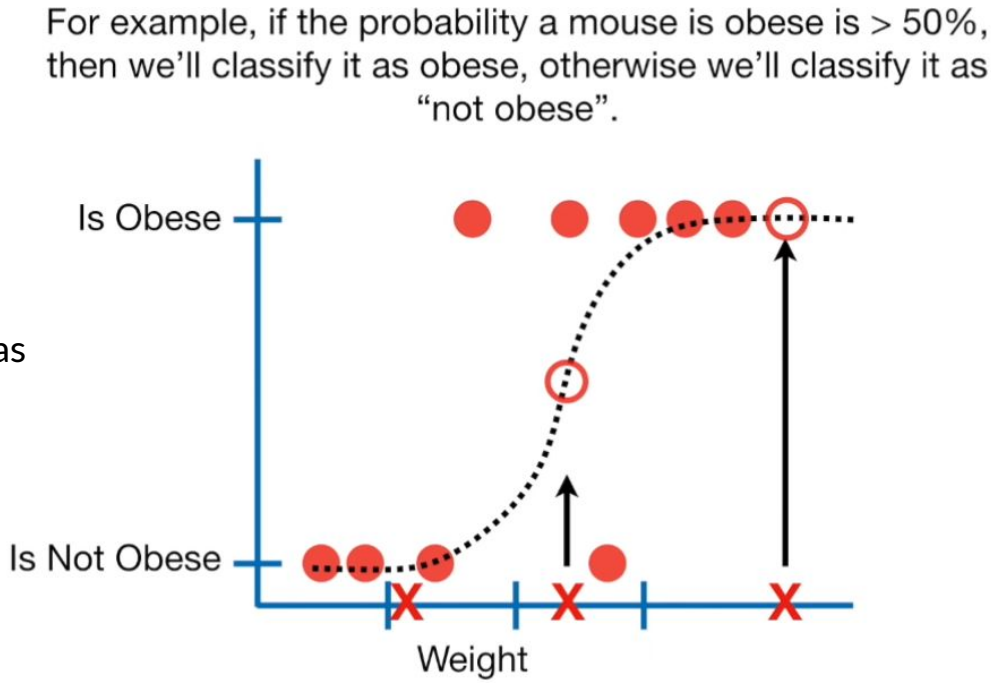
Entrada: Dos dimensiones, datos
Salida: Un valor estimado (número)

¿Correlación?



Logistic Regression

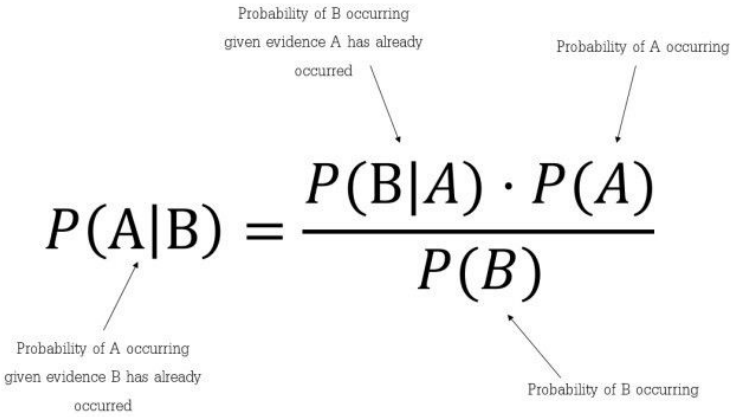
Entrada: Una dimensión, dos categorías
Salida: Probabilidad **(clasificación)**



Naive Bayes

Algoritmo probabilístico. Teoría → Conocimiento parcial, proceso estocástico (no determinista)

Key: variables → tabla de frecuencia → probabilidades e incertidumbre



Naive Bayes

Saturdays.AI
Donostia

Dataset

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Tabla de Frecuencias

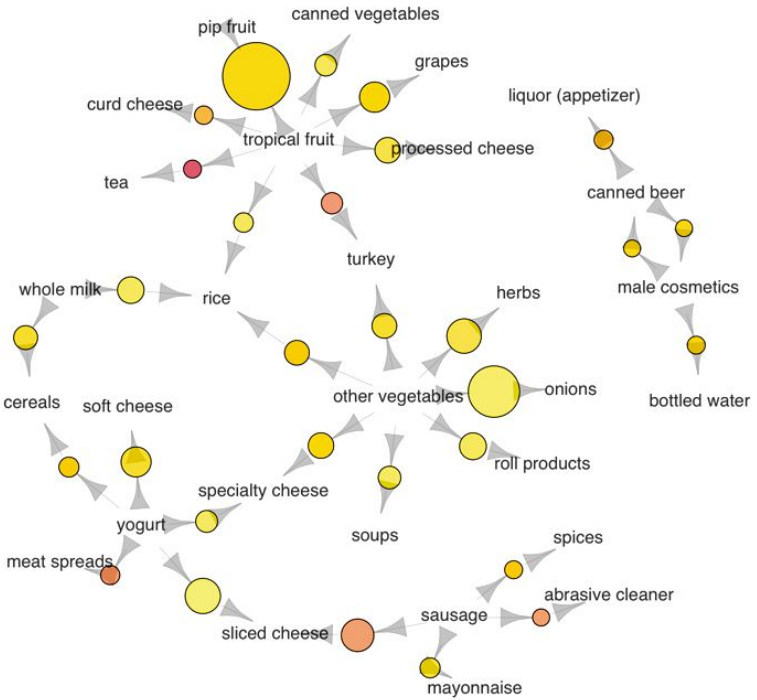
Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Probabilidades

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

→ **Laplace trick**

Association Rules



$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

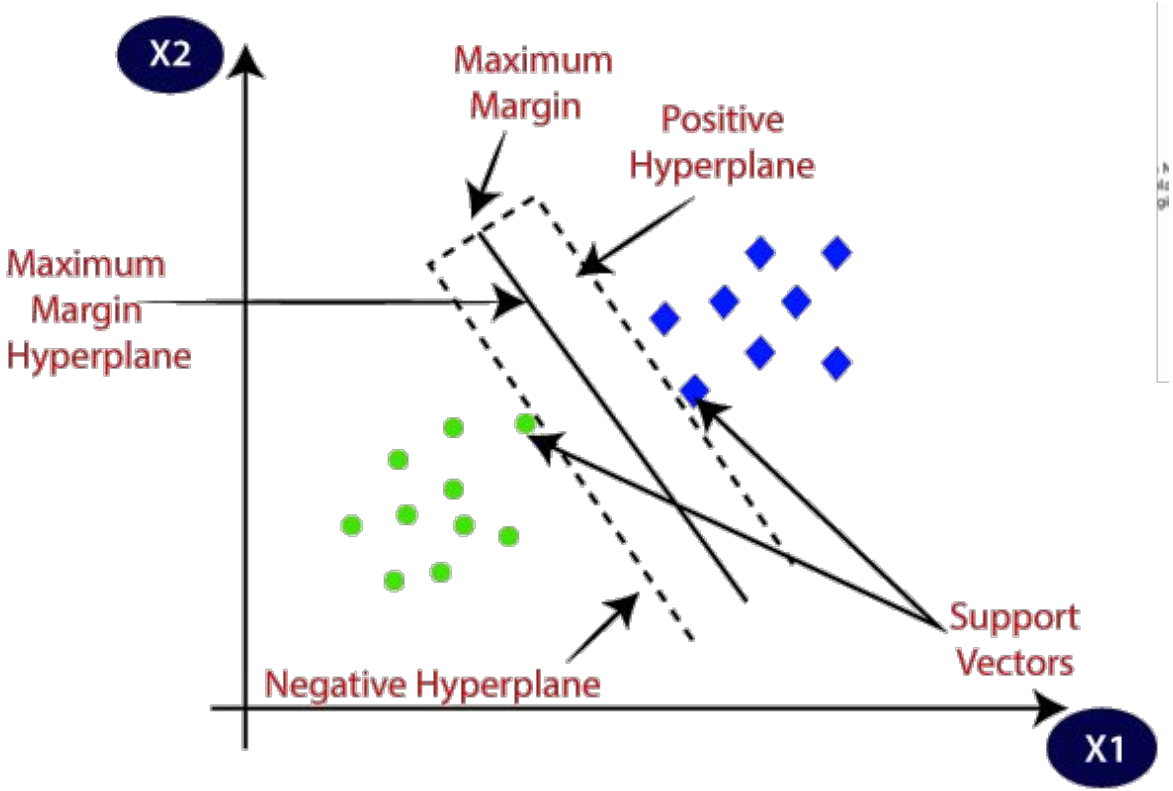
$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Association Rules

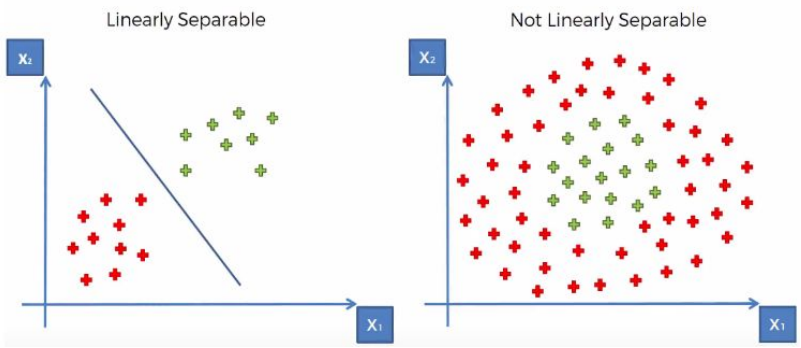
Saturdays.AI
Donostia

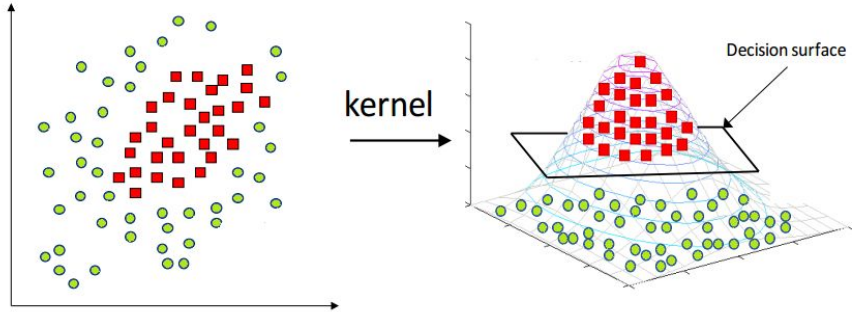
Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

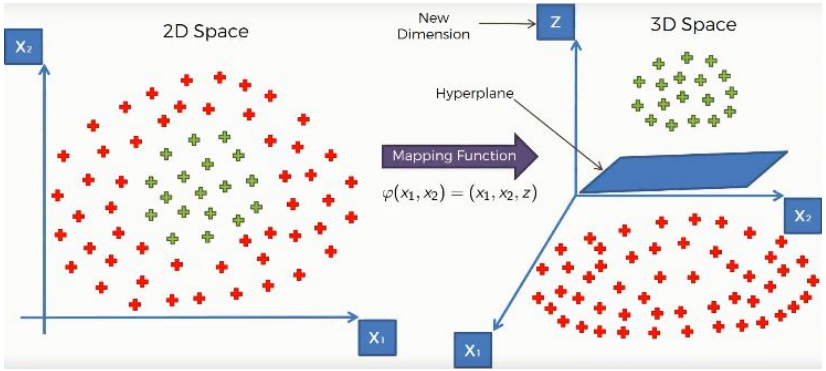
Support Vector Machine



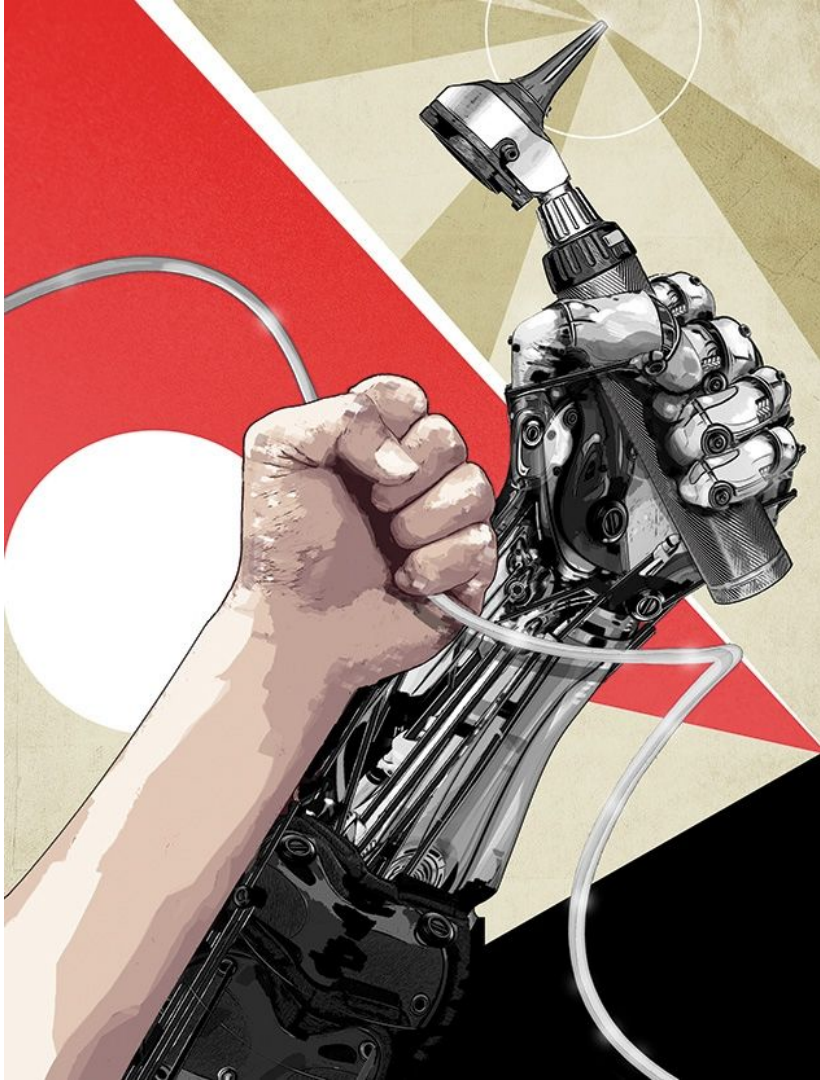
Support Vector Machine



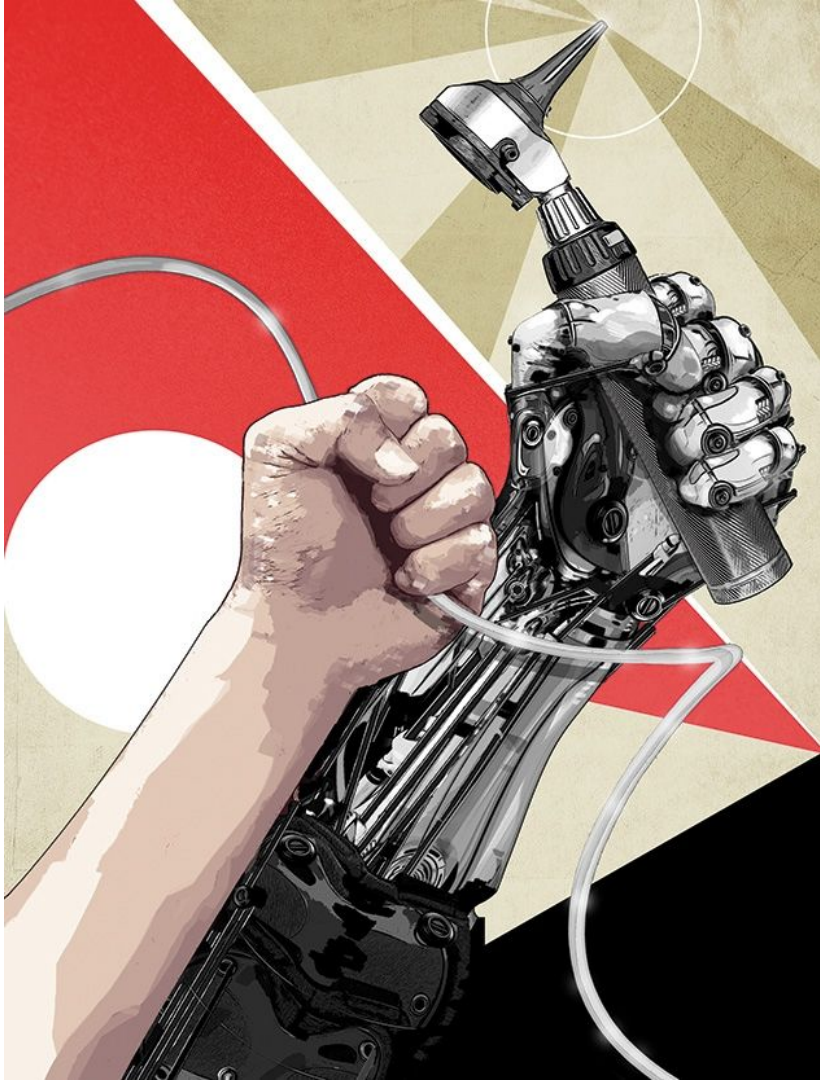




Practice - Supervised Learning!



Challenge - Supervised Learning!



Bibliografía

/1./ /Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow/

/2./ /Fast.AI - Introduction to Machine Learning for Coders/

/3./ /MLCourse.AI/

/4./ /DeltaAnalytics/

/5./ /The Hundred-page Machine Learning Book/

/6./ /Machine Learning for Humans (Vishal Maini)/

/7./ /Datacamp/

/8./ /DataQuest/

Partners

Saturdays.AI Donostia

Agradecemos a nuestros partners por confiar en nosotros para facilitar la formación en IA de cara a la 4ª Revolución Industrial.

IMPACT HUB

Bizkaia foru aldundia diputación foral

spri ENPRESA DIGITALA

eureka! zientzia museoa kutxa FUNDAZIOA

Deusto Universidad de Deusto

kutxa kultur kutxa FUNDAZIOA

eman ta zabal zazu UPV EHU

Celera Aceleradora de personas

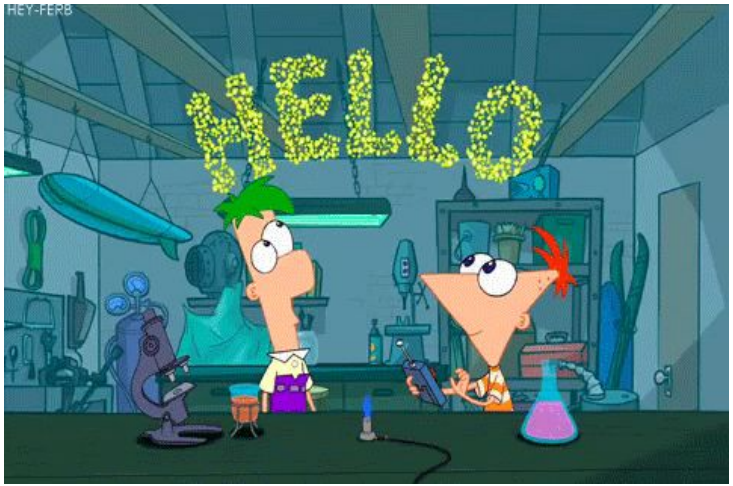
NVIDIA INCEPTION PROGRAM

aws

4YFN

Saturdays.AI
Donostia

WELCOME!



Saturdays.AI

This model fits me
95% of the time

CAMP TECNOLÓGICO

GLOBAL SHAPERS COMMUNITY BILBAO

www.saturdays.ai

donostia.saturdays.ai

donostia@saturdays.ai